Big Data Driven Similarity Based U-Model for Online Social Networks

Jingjing Wang*, Chunxiao Jiang†, Sanghai Guan*, Lei Xu* and Yong Ren*
*Department of Electronic Engineering and †Tsinghua Space Center, Tsinghua University, Beijing, 100084, China
Email: chinaeephd@gmail.com, jchx@tsinghua.edu.cn, guansanghai@gmail.com, xu_l04@163.com, reny@tsinghua.edu.cn

***Abstract*—The proliferation of information technologies results in a complex network evolution of online social networks. Traditional model driven aided description cannot be appropriate for the dynamic evolution of social networks. However, in this paper, relying on the big data collected from a range of real-world online social networks, we try to explore the underlying evolution for online social networks. Firstly, we define a pair of big data driven similarity based utility models (U-models), i.e. the undirected U-model as well as the directed U-model, which can effectively reflect the statistical characteristics of online social networks. Secondly, we analyze the small-world property, scale-free property and high clustering coefficient property of our proposed U-models which consider nodes' similarity, popularity and asymmetry in a network. Finally, relying on three real-world big datasets, i.e. Sina Weibo, Tencent Weibo and Twitter, sufficient experiments show that the U-models outperform the traditional models in portraying the evolution statistical characteristic of online social networks.**

***Index Terms*—Big data driven, network evolution, similarity based utility, complex network.**

I. Introduction

With the development of information technologies and the expansion of network scale, the amount of data in online social networks increases exponentially [1]. Massive social data and increasingly mature data mining technologies make it possible for providing compelling applications and enhancing user experiences [2], [3]. Hence, in this 'big data' era, establishing an appropriate network evolving model is beneficial in terms of reasonable resource allocation and of accurate users' behavior analysis [4].

In the literatures, the big data driven online social network modeling has attracted much attention [5], [6]. Complex network theory was proposed in order to portray such a network with a large number of nodes and interactive connectivity. Barabási *et al*. [7] pointed out that data-based mathematical models of complex network were developing into the network science. Specifically, Watts *et al*. [8] proposed the WS 'small-world' model, which had a short average path length and a high clustering coefficient. In [9], Barabási *et al.* discovered the scale-free property from real-world social networks and presented the BA network model, where the node degree distribution always met the power-law distribution. Moreover, he indicated that the edges' connectivity possessed the 'preferential attachment' characteristic. Then, there were a variety of relevant modified models relying on the above two models [10]–[13]. Li *et al.* [14] proposed a novel network evolving model in terms of a new concept of 'local-world connectivity'. They pointed out that this local-world evolving model was capable of both maintaining the robustness of scale-free networks as well as of improving the network's reliance against intentional attacks. Moreover, Gu *et al.* [15] presented a new network growth rule which considered the node addition, node deleting and the 'local-world connectivity'. In [16], Li *et al.* proposed a community structure aided network evolving model verified by theoretical analysis and numerical simulations. Furthermore, Cao *et al.* [17] provided a neighborhood connectivity based network evolving model, which can be used to enhance the evolving mechanism of real-world complex networks, such as online social networks, vehicular networks, etc. Papadopoulos *et al.* [18] developed a evolving framework, where new connections followed the optimal trade-off between popularity and similarity rather than simply connecting to popular nodes.

However, numerous real-world social networks simultaneously have three properties, i.e. the small-world property, the scale-free property and the high clustering coefficient property, while most network evolving models mentioned above are just characterized by part of them [19]. Furthermore, the proposed network evolving models only focus their attention on undirected networks, but the real-world social networks are invariably directed resulting in ubiquitous asymmetry [20].

Inspired by the above open challenges, in this paper, relying on the real-world big datasets, we proposed a pair of big data driven similarity based utility models (U-model), namely undirected U-model and directed U-model, for online social networks in order to explore their evolution characteristics. Moreover, our directed U-model is beneficial of constructing the ubiquitous asymmetry. Our original contributions are summarized as follows:

- A similarity-based U-model is proposed by considering the interest similarity between nodes when constructing the utility function.
- Given the asymmetry in directed networks, the directed U-model is elaborated relying on the asymmetry utility function as well as the asymmetry preferential attachment.
- Sufficient experiments are conducted based on three real-world big datasets including Sina Weibo, Tencent Weibo and Twitter. Moreover, we verify the feasibility and the reliability of our proposed model which matches the real-world networks.

978-1-5090-5019-2/17/$31.00 ©2017 IEEE

The rest of the paper is organized as follows. In Section II, we establish a pair of similarity based utility models and define their utility functions. In Section III, a range of experiments and simulations are conducted in order to verify the superior performance of our proposed models relying on three real-world big datasets, followed by our conclusions and future work in Section IV.

II. System Model

A. Definition of U-model

In this section, we first define the utility based evolving network model. In our proposed U-model, when a new node t joins in the growing network, it will connect to one of the existed nodes with a probability, which depends on the utility of the connection. Here the utility can be formulated as:

$$u(i,t) = b(i,t) - c(i,t) + \rho_t, \tag{1}$$

where $u(i,t)$ represents the utility that node t will achieve if it connects to node i. Moreover, $b(i,t)$ denotes the total benefit that node t receives, and $c(i,t)$ is the cost that node t needs to pay for the connection. The perturbation factor of node t is given by ρ_t, which describes the irrational behavior.

Furthermore, the utility is related to nodes' degree, similarity, betweenness centrality, asymmetry and other factors. Hence, we define the utility function in another way to study both the network evolution process and their impact. The utility based on parameters can be given by:

$$u(i,t) = \prod_n F_n(i,t)^{\alpha_n}, \tag{2}$$

where $F_n(i,t)$ denotes different parameters and α_n represents the exponents of different parameters. The n-th parameter's impact on the utility is enlarged with the increasing of α_n.

As a result, we propose a network construction algorithm for the U-model as follows:

1) Network growth: Initially, a fully connected network G with m_0 nodes is formed. At each step, a new node t joins in the network and selects m existed nodes to connect.
2) Preferential attachment: The probability of p_{it} that benchmarks the possibility of a new node t connecting to an existed node i, is proportional to the utility $u(i,t)$, which can be described as:

$$p_{it} = \frac{u(i,t)}{\sum_{j \in G} u(j,t)}, \tag{3}$$

where G represents nodes' set of the network.

Here, we present two toy cases in order to further explain the U-model. Specifically, when $u(i,t) = k_i$ (the degree of existed node i), the U-model corresponds to the BA model. When $u(i,t) = c$ (a constant number), the U-model describes a homogeneous random network. In the following, we will introduce a similarity-based U-model which is capable of well matching online social networks.

B. Undirected Similarity-Based U-model

As discussed in Section I, BA model only takes the popularity of nodes into account, which results in a low clustering coefficient. Thus, the BA model may not exactly describe the clustering feature of online social networks. In online social networks, similarity between two nodes becomes the main cause of the clustering. For example, people prefer to interact with others who have similar interests with them. Hence, the popularity and the similarity are two important factors for constructing the utility. In our model, we define the similarity-based utility function as:

TABLE I
Shorthand symbols

Symbol	Description
N	Number of nodes
E	Number of edges
$\langle k \rangle$	Average degree
L	Average path length
r	Exponent of power-law distribution
C	Clustering coefficient

TABLE II
Average path length and clustering coefficient of undirected similarity-based U-model*

	$\alpha = 0$ (BA)	$\alpha = 1$	$\alpha = 2$	$\alpha = 3$	$\alpha = 4$
L	5	5.4	8.4	10	11
C	0.001	0.24	0.40	0.52	0.54

* $N = 10{,}000$, $m_0 = 3$, $m = 3$.

$$u(i,t) = s_{it}^{\alpha} k_i \qquad (\alpha > 0), \tag{4}$$

where $s(i,t)$ denotes the similarity between node i and t, and we have $s(i,t) = s(t,i)$. Moreover, α represents the weighting factor of similarity and k_i is the degree of node i. If $\alpha = 0$, utility depends entirely on nodes' popularity and the network degenerates to a BA network. Then, the connection probability defined in Eq. (3) can be rewritten as:

$$p_{it} = \frac{s_{it}^{\alpha} k_i}{\sum_{j \in G} s_{jt}^{\alpha_1} k_j}. \tag{5}$$

In this paper, we use a random number between 0 and 1, namely u_i, to represent node i's interest. The similarity of interests between node i and node t is formulated as:

$$s(i,t) = \frac{1}{|u_i - u_t|}. \tag{6}$$

Hence, we construct an undirected similarity-based network having total $N = 10000$ nodes according to the above definitions with initial variables $m_0 = 3$ and $m = 3$. Some main characteristics of the network are described in Table II and Fig. 1. Moreover, some shorthand symbols in this paper are summarized in Table I.

Table II indicates that the clustering coefficient of the network depends on the weight of similarity in utility function. As α increasing, the clustering coefficient of the network increases as well because similar nodes get together. From

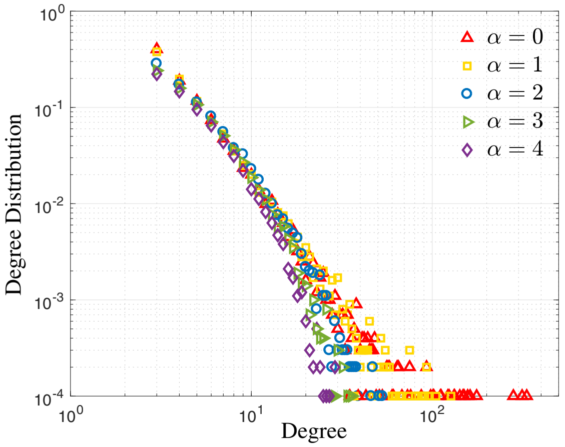

Fig. 1. Degree distribution of undirected similarity-based U-model. ($N = 10,000$, $m_0 = 3$, $m = 3$)

Fig. 1, we can find that the degree distribution of U-model satisfies scale-free property. The curves inward as α getting larger for similarity influence the connections and popular nodes have less attraction. It also results in the increasing of average path length of the network because popular nodes contribute to shortening it.

C. Directed Similarity-Based U-model

Numerous real-world networks are characterized by the directed structure and asymmetry. In a directed network, the in-degree and the out-degree of a node are not equal. As for online social networks, such as the Twitter for example, the lists of a user's 'following' and 'follower' are generally different. This phenomenon can be viewed as a kind of asymmetry, which is of vital importance in the evolution of the network. In this subsection, we will propose a directed U-model based on asymmetry, popularity and similarity.

In our model, the in-degree of a node can be considered as its 'popularity', which is denoted by $k_{in}(t)$. Hence, Eq. 4 can be rewritten as:

$$u(i,t) = s(i,t)^{\alpha} k_{in}(i) \quad (\alpha > 0). \tag{7}$$

Then, the probability of the connection from a new node t to the existed node i can be formulated as:

$$p_{it} = \frac{s(i,t)^{\alpha} k_{in}(i)}{\sum_{j \in G} s(i,t)^{\alpha} k_{in}(i)}. \tag{8}$$

When the new node t connects to node i, whether node i establishes the inverted connection depends on the asymmetry utility. We assume that people in online social networks are not absolutely rational. For example, a user may follow someone who has followed him just by curiosity or casualness. Hence, we consider the irrational effect into the asymmetry function utility, i.e.

$$u_a(i,t) = s(i,t) r(i)^{\beta} \quad (\beta > 0), \tag{9}$$

where $u_a(i,t)$ represents the asymmetry utility of the existed node i. And $u(i,t)$ is not always equal to $u_a(i,t)$. Moreover, $r(i)$ denotes the irrational factor of node i and β is the weighting factor of the irrational factor.

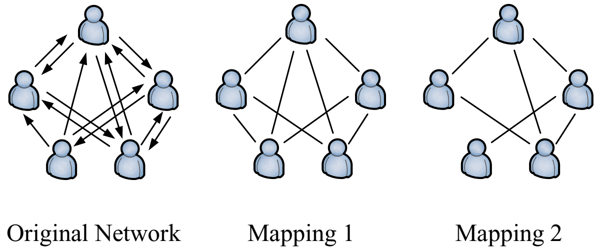

Fig. 2. Two mappings from directed network to undirected network.

TABLE III
AVERAGE PATH LENGTH AND CLUSTERING COEFFICIENT OF DIRECTED SIMILARITY-BASED U-MODEL UNDER TWO MAPPINGS*

Mapping		$\beta = 0$	$\beta = 1$	$\beta = 2$	$\beta = 3$	$\beta = 4$
1	L	3.6	3.6	3.6	3.6	3.6
	C	0.08	0.08	0.08	0.08	0.08
2	L	5.2	5.1	4.9	4.8	4.7
	C	0.18	0.17	0.15	0.13	0.12

* $N = 10,000$, $m_0 = 6$, $m = 6$, $\alpha = 1$, $p = 0.5$.

As to construct this directed network, at each step, the new node t selects m connections relying on p_{it}. Then q nodes establish the inverted connection to node t. Hence, the bidirectional connections are formed. Here, we have $p = q/m$, where p is the ratio of asymmetry connection. The asymmetry connection probability can be given by:

$$p_{ti} = \frac{s(i,t) r(i)^{\beta}}{\sum_{j \in M(t)} s(i,t) r(i)^{\beta}}, \tag{10}$$

where $M(t)$ is the set of existed nodes connected by node t. Hence, relying on the above definitions, we construct a directed network with $N = 10000$, $m_0 = 6$, $m = 6$, $\alpha = 1$ and $p = 0.5$.

For the feasibility of analysis, the directed network can be mapped into an undirected network. In the following, we define two mappings named mapping 1 and mapping 2, which are illustrated in Fig. 2. The influence of the irrational factor on some networks features under two mappings are listed in Table III.

As shown in Table III, the weighting factor β has no effect in the context of mapping 1. In contrast, under mapping 2, with the increasing of β, the impact of similarity on the network's connection becomes weak, where the clustering coefficient is small and the average path length is large. Moreover, the clustering coefficient in the context of mapping 2 is larger than that in mapping 1, for the former represents a stronger relationship between two nodes in networks, while it also results in a longer average path. Therefore, we will focus our attention on the context of mapping 2 in the following.

TABLE IV
AVERAGE PATH LENGTH AND CLUSTERING COEFFICIENT OF RATIONAL DIRECTED SIMILARITY-BASED U-MODEL UNDER MAPPING 2*

	$\alpha=0$	$\alpha=1$	$\alpha=2$	$\alpha=3$	$\alpha=4$
L	3.9	5.1	7.3	8.2	9
C	0.02	0.19	0.43	0.48	0.52

* $\beta=0$, $N=10,000$, $m_0=6$, $m=6$, $p=0.5$.

TABLE V
AVERAGE PATH LENGTH AND CLUSTERING COEFFICIENT OF IRRATIONAL DIRECTED SIMILARITY-BASED U-MODEL UNDER MAPPING 2*

	$\alpha=0$	$\alpha=1$	$\alpha=2$	$\alpha=3$	$\alpha=4$
L	3.9	4.8	6.5	7.8	9
C	0.02	0.11	0.34	0.44	0.49

* $\beta=4$, $N=10,000$, $m_0=6$, $m=6$, $p=0.5$.

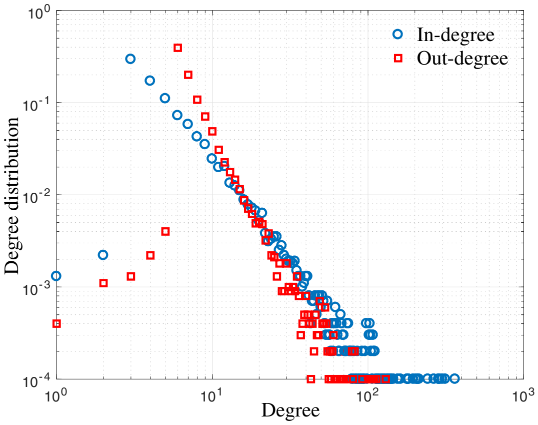

Fig. 3. Degree distribution of directed similarity-based U-model under the assumption of rationality. ($\beta=0$, $N=10,000$, $m_0=6$, $m=6$, $\alpha=1$, $p=0.5$).

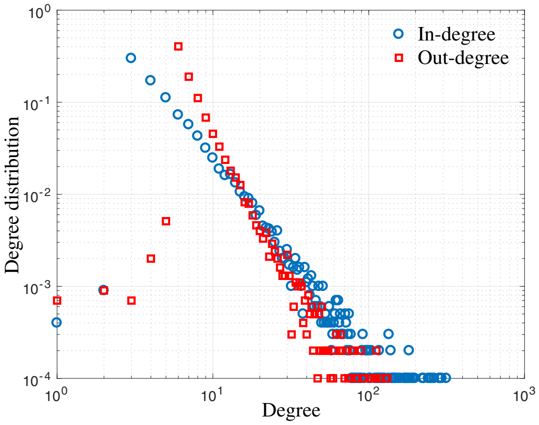

Fig. 4. Degree distribution of directed similarity-based U-model under the assumption of irrationality. ($\beta=4$, $N=10,000$, $m_0=6$, $m=6$, $\alpha=1$, $p=0.5$).

The characteristics of the network under rational and irrational conditions are summarized in Table IV, Table V, Fig. 3 and Fig. 4, where let $\beta=0$ and $\beta=4$, respectively. From Fig. 3 and Fig. 4, we can find that although the different irrational factor leads to different degrees of nodes, the degree distribution shows no significant difference from a macroscopic perspective. The power exponents r in power-law distribution of in-degree and out-degree are about 2.5 and 2, respectively. As a result, our directed similarity-based U-model generates asymmetric in-degree and out-degree distribution which well matches the real-world online social networks.

III. EXPERIMENTS

In this section, relying on three real-world datasets, we analyze the performance of our proposed U-models compared with the traditional BA model.

A. Three Datasets of Online Social Networks

In the following, we first introduce three real-world datasets of online social networks, i.e. *Twitter* and two popular Chinese microblogs named *Sina Weibo* and *Tencent Weibo*.

1) Sina Weibo: Sina Weibo[1] is a Chinese microblogging website and one of the most popular sites in China. By the third quarter of 2015, Sina Weibo has 222 million subscribers and 100 million daily users. About 100 million messages are posted each day on it. We get the dataset from *WISE 2012 Challenge*[2]. This dataset contains over five million nodes and over 330 million edges.

2) Tencent Weibo: Tencent Weibo[3] is a Chinese microblogging website launched by Tencent in April 2010 and has more than 100 million subscribers. We obtain this dataset from *KDD Cup 2012*[4], which has about two million nodes and about 50 million edges.

3) Twitter: Twitter[5] is an online news and social networking service where users post and interact with messages restricted to 140 characters. In 2016, Twitter had more than 319 million monthly active users. We acquire the Twitter dataset containing about five million nodes and eight million edges from *Aminer*[6].

From each dataset, we select 10000 nodes and and the connections among them for experiment.

B. Undirected Network

In our experiment, we map the real-world datasets under mapping 1 to get undirected networks. Then we adjust the parameters of both undirected U-model and BA model to construct undirected networks and fit the real-world datasets. The comparison of statistical characteristics are shown in Table VI, Fig. 5, Fig. 6 and Fig. 7. We can conclude that undirected

[1] http://weibo.com

[2] http://www.wise2012.cs.ucy.ac.cy/challenge.html

[3] http://t.qq.com

[4] https://www.kaggle.com/c/kddcup2012-track1#description

[5] https://twitter.com

[6] https://aminer.org/data-sna

TABLE VI
COMPARING OF UNDIRECTED SIMILARITY-BASED U-MODEL, BA MODEL AND REAL-WORLD DATASETS UNDER MAPPING 1

Network	N	E	$\langle k \rangle$	L	r	C
Sina Weibo	10,000	97,103	20	4.0	2.3	0.12
U-model	10,000	98,266	20	4.0	2.4	0.12
BA model	10,000	98,334	20	3.0	2.5	0.01
Tencent Weibo	10,000	414,314	83	2.6	2.2	0.19
U-model	10,000	419,007	84	2.6	2.4	0.19
BA model	10,000	419,095	84	2.4	2.5	0.03
Twitter	10,000	41,527	8	3.8	1.9	0.05
U-model	10,000	39,840	8	4.1	1.9	0.05
BA model	10,000	39,990	8	3.9	2.0	0.006

U-model shows much better performance than BA model in describing online social networks, especially the clustering coefficient of the network. It satisfies the small-world property, scale-free property and high clustering coefficient at the same time on the macro level.

C. Directed Network

Because most network models, including BA model, only focus on undirected networks, we will only compare the datasets with our directed U-model in this subsection. In our experiment, we adjust the parameters of directed similarity-based U-model to fit datasets of Sina Weibo and Tencent Weibo. Then we map the results into undirected networks under mapping 2 for analysis.

As can be seen from Table VII, directed U-model satisfies the characteristics of the real-world directed network under mapping 2 accurately. From Fig. 8 and Fig. 9, we can also find that the directed U-model describes the characteristics of the real-world network on degree distribution, distance distribution and clustering coefficient precisely. So we can conclude that U-model effectively describes the evolution of real-world networks on the macro level.

IV. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed two kinds of similarity-based U-model, i.e. the undirected U-model and the directed U-model. Our proposed U-models satisfied the properties of small-world, scale-free and high clustering coefficient simultaneously. Moreover, our directed U-model reveals the asymmetry in social networks. In our experiments, U-model shows a better performance than the traditional BA model, and it has unique advantages in describing directed networks. In future work, we intend to improve the mathematical deduction of U-model and figure out its statical characteristics analytically. Moreover, We will explore the benefit of our proposed U-model applied in information diffusion process for online social networks.

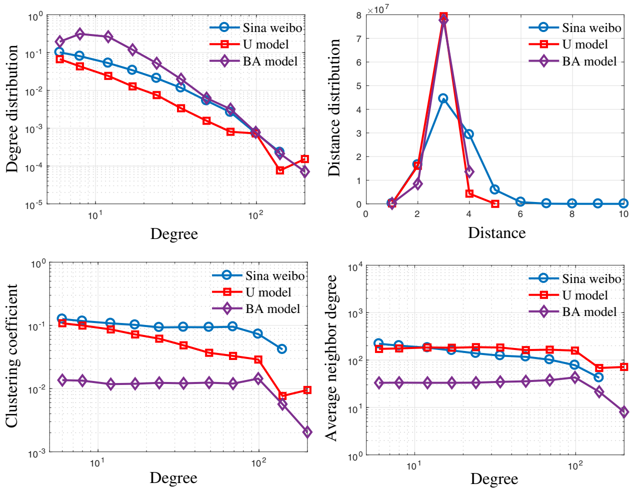

Fig. 5. Experiment about Sina Weibo under mapping 1.

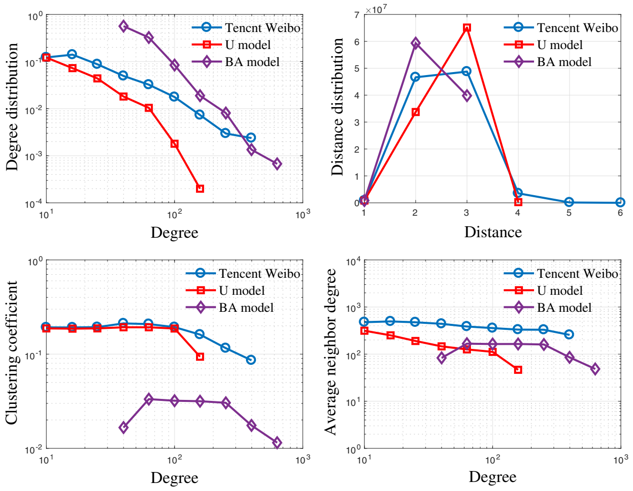

Fig. 6. Experiment about Tencent Weibo under mapping 1.

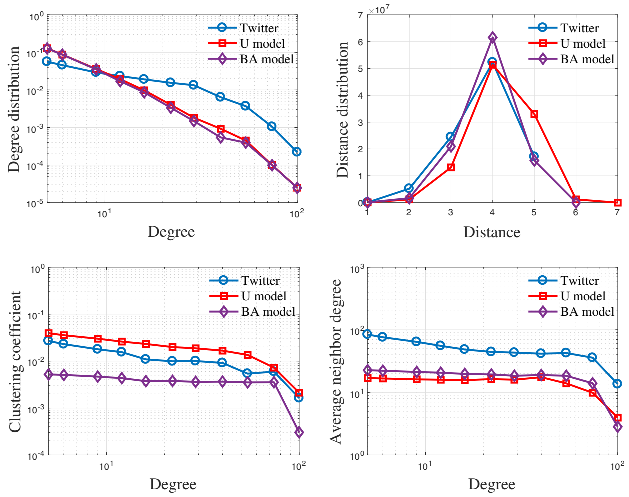

Fig. 7. Experiment about Twitter under mapping 1.

TABLE VII
COMPARING OF DIRECTED SIMILARITY-BASED U-MODEL AND REAL-WORLD DATASETS UNDER MAPPING 2

Network	N	E	$\langle k \rangle$	L	r	C
Sina Weibo	10,000	19,077	4	5.0	2.0	0.12
U-model	10,000	19,980	4	5.0	2.0	0.12
Tencent Weibo	10,000	27,412	6	4.4	2.0	0.32
U-model	10,000	29,828	6	5.0	2.5	0.34

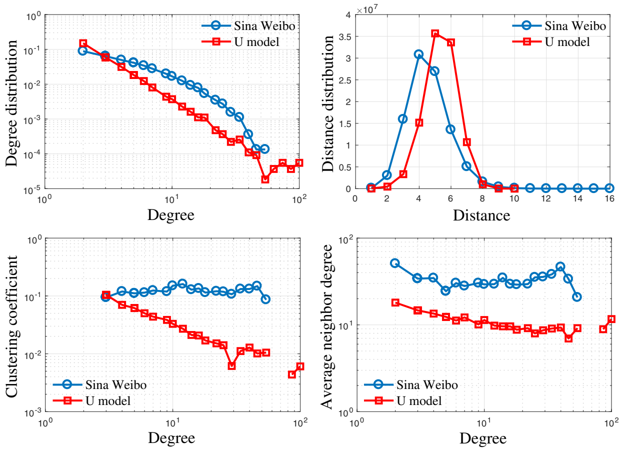

Fig. 8. Experiment about Sina Weibo under mapping 2.

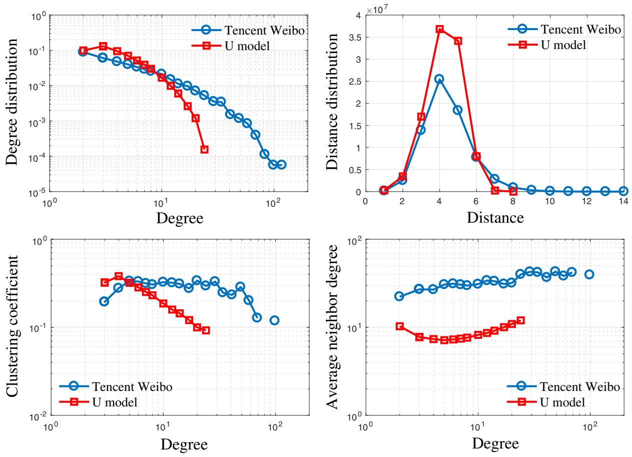

Fig. 9. Experiment about Tencent Weibo under mapping 2.

ACKNOWLEDGEMENT

This research was supported by NSFC China under projects 61371079 and 61471025.

REFERENCES

[1] W. Tan, M. B. Blake, I. Saleh, and S. Dustdar, "Social-network-sourced big data analytics," *IEEE Internet Computing*, vol. 17, no. 5, pp. 62–69, 2013.

[2] J. Wang, C. Jiang, T. Q. Quek, X. Wang, and Y. Ren, "The value strength aided information diffusion in socially-aware mobile networks," *IEEE Access*, vol. 4, pp. 3907–3919, 2016.

[3] C. Jiang, Y. Chen, Y. Ren, and K. R. Liu, "Maximizing network capacity with optimal source selection: A network science perspective," *IEEE Signal Processing Letters*, vol. 22, no. 7, pp. 938–942, 2015.

[4] S. Kulcu, E. Dogdu, and A. M. Ozbayoglu, "A survey on semantic web and big data technologies for social network analysis," in *2016 IEEE International Conference on Big Data*. IEEE, 2016, pp. 1768–1777.

[5] J. Wang, C. Jiang, T. Q. Quek, and Y. Ren, "The value strength aided information diffusion in online social networks," in *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Washington, DC, USA, Dec. 2016, pp. 1–6.

[6] C. Jiang, Y. Chen, and K. R. Liu, "Graphical evolutionary game for information diffusion over social networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 4, pp. 524–536, 2014.

[7] A.-L. Barabási, "The network takeover," *Nature Physics*, vol. 8, no. 1, pp. 14–16, 2011.

[8] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[9] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[10] D. A. Fell and A. Wagner, "The small world of metabolism," *Nature biotechnology*, vol. 18, no. 11, pp. 1121–1122, 2000.

[11] R. Cohen and S. Havlin, "Scale-free networks are ultrasmall," *Physical review letters*, vol. 90, no. 5, p. 058701, 2003.

[12] L. A. Adamic and B. A. Huberman, "Power-law distribution of the world wide web," *science*, vol. 287, no. 5461, p. 2115, 2000.

[13] A. Fronczak, P. Fronczak, and J. A. Hołyst, "Mean-field theory for clustering coefficients in barabási-albert networks," *Physical Review E*, vol. 68, no. 4, p. 046126, 2003.

[14] X. Li and G. Chen, "A local-world evolving network model," *Physica A: Statistical Mechanics and its Applications*, vol. 328, no. 1, pp. 274–286, 2003.

[15] Y. Gu and J. Sun, "A local-world node deleting evolving network model," *Physics Letters A*, vol. 372, no. 25, pp. 4564–4568, 2008.

[16] C. Li and P. K. Maini, "An evolving network model with community structure," *Journal of Physics A: Mathematical and General*, vol. 38, no. 45, p. 9741, 2005.

[17] Y. J. Cao, G. Z. Wang, Q. Y. Jiang, and Z. X. Han, "A neighbourhood evolving network model," *Physics Letters A*, vol. 349, no. 6, pp. 462–466, 2006.

[18] F. Papadopoulos, M. Kitsak, M. Á. Serrano, M. Boguná, and D. Krioukov, "Popularity versus similarity in growing networks," *Nature*, vol. 489, no. 7417, pp. 537–540, 2012.

[19] C. Jiang, Y. Chen, and K. R. Liu, "Evolutionary dynamics of information diffusion over social networks," *IEEE Transactions on Signal Processing*, vol. 62, no. 17, pp. 4573–4586, 2014.

[20] L. Jin, Y. Chen, T. Wang, P. Hui, and A. V. Vasilakos, "Understanding user behavior in online social networks: A survey," *IEEE Communications Magazine*, vol. 51, no. 9, pp. 144–150, 2013.